doi.org/10.54473/IJTRET.2024.8105
International Journal of Trendy Research in Engineering and Technology
Volume 8 Issue 1 February 2024
ISSN No.2582-0958

AN INTEGRATED MACHINE LEARNING FRAMEWORK FOR ACCURATE CARDIOVASCULAR DISEASE PREDICTION

Rakesh V, Sathya narayanan G,
Department of Electronics and Communication Engineering, Easwari Engineering College, Chennai
Rakesh.v0717@gmail.com , sn297416@gmail.com

Received 25 Feb 2024 Received in revised form 04 March 2024 Accepted 08 March 2024
Available online 12 March 2024

ABSTRACT

This paper introduces an advanced real-time system designed to predict cardiovascular diseases with integrated machine learning. Cardiovascular diseases, with the highest global mortality rate, have become increasingly prevalent, straining healthcare systems worldwide. These diseases, driven by factors such as high blood pressure, stress, age, gender, and cholesterol levels, have prompted numerous early diagnosis approaches, but their accuracy requires refinement due to the critical nature of cardiovascular diseases. This paper presents the DLCDD (Deep Learning based Cardiovascular Disease Diagnosis) framework, specifically addressing data-related challenges such as missing values and imbalances. The mean replacement technique is employed for handling missing values, while the Synthetic Minority Over-sampling Technique (SMOTE) is utilized to address imbalances in the dataset. In essence, DLCDD represents a significant advance in precise cardiovascular disease prediction, uniting deep learning with cutting-edge data processing and feature selection methods, addressing critical diagnostic challenges.

Keywords— Integrated Machine Learning, Cardiovascular Disease, Electronic Health Records, BILSTM, SMOT, Prediction, Healthcare Analytics.

I. INTRODUCTION

In the context of an alarming surge in cardiovascular disease-related mortality rates worldwide, addressing this pervasive health issue has gained paramount significance within the global healthcare landscape. Cardiovascular diseases, encompassing a spectrum of conditions such as hypertension, stress-induced ailments, and atherosclerosis, are intricate disorders influenced by a multitude of factors including age, gender, cholesterol levels, and genetic predisposition. Given the life-threatening implications associated with cardiovascular diseases, the pursuit of advanced diagnostic methods offering superior accuracy has become imperative.

While considerable Advancements have been achieved in the development of existing diagnostic techniques for early detection, the inherent complexity of cardiovascular diseases mandates an ongoing refinement effort to enhance their precision. The consequences of false negatives or misdiagnoses are profound, leading to delayed intervention and potentially fatal outcomes for afflicted individuals. In this context, the introduction of the DLCDD (Deep Learning based Cardiovascular Disease Diagnosis) framework signifies a significant leap towards improving the accuracy and efficacy of cardiovascular disease prediction.

The DLCDD framework constitutes a multifaceted approach addressing critical aspects of cardiovascular disease diagnosis. It commences with a meticulous treatment of missing data through the mean replacement technique, ensuring the dataset used for analysis is both complete and reliable. By mitigating the impact of missing data, the framework lays a robust foundation for subsequent stages of analysis.

Another pervasive challenge in cardiovascular disease prediction models is data imbalance. To address this issue, the DLCDD framework utilizes the Synthetic Minority Over-sampling Technique (SMOTE), a technique that produces artificial data points for the minority class, thereby attaining a more equitable dataset [1]. This step guarantees that the model remains unbiased toward the majority class and can effectively identify individuals at risk, irrespective of the prevalence of the condition.

A pivotal feature of the DLCDD framework is its judicious use of Feature Importance techniques. These techniques facilitate the identification of the most influential variables

within the dataset [2]. By selecting the most relevant features, the framework optimizes the accuracy and efficiency of the predictive model. This strategic approach not only heightens the precision of cardiovascular disease prediction but also streamlines the diagnostic process, this approach has the potential to facilitate earlier interventions, ultimately enhancing patient outcomes.

The global healthcare community confronts the pressing challenge of combatting cardiovascular diseases, innovative solutions such as the DLCDD framework offer a ray of hope. By addressing missing data, tackling data imbalance, and optimizing feature selection, this framework represents a comprehensive strategy to significantly enhance the accuracy and effectiveness of cardiovascular disease prediction. As we endeavour to reduce the burden of cardiovascular diseases on healthcare systems and improve patient well-being, embracing advanced diagnostic approaches is not merely a necessity; it is an indispensable imperative for a healthier and more resilient future.

II. LITERATURE REVIEW

In this section, we provide a concise overview of the relevant literature concerning on predicting cardiovascular diseases with integrated machine learning.

Coffee is one of the most popular beverages worldwide. The positive impact of coffee on health is widely acknowledged in numerous studies. Nevertheless, increasing apprehensions surround the potential negative influence of consuming coffee on cardiovascular disease (CVD) [3]. This concern arises from the perceived exacerbating effects on the cardiovascular system attributed to various compounds present in coffee.

Ghorashi et al [4] have made notable contributions to the realm of medical informatics, particularly in the utilization of deep learning models for the prediction and diagnosis of cardiovascular diseases (CVD). Their investigation involves the implementation of a regression analysis based on deep learning techniques, utilizing a dataset comprising 2621 patient medical files sourced from hospitals in the United Arab Emirates (UAE). The dataset encompasses essential variables such as age, symptoms, and CVD information to enable the early prediction of CVDs. The proposed long short-term memory-based deep neural network exhibits enhanced accuracy, particularly when considering diseases in pairs or combinations due to overlapping symptoms.The accuracy in predicting coronary heart disease is 71.5%, which rises to 88.9% when considering a combination of symptoms such as Dyspnea, cyanosis, hemoptysis, fatigue, weakness, chest pain, and chest discomfort are indicative symptoms that may warrant medical attention. When Fever is included, the method achieves an accuracy of 91%. These results underscore the effectiveness of the proposed approach across various evaluation benchmarks, emphasizing its potential in advancing the accuracy of CVD prediction and early intervention strategies.

Barbara Riegel et al [5]delve into the significance of self-care in chronic illness prevention and management. They emphasize the naturalistic decision-making processes involved, which encompass maintenance, monitoring, and active management. The authors emphasize the significance of self-care within the American Heart Association's mission, particularly in the context of cardiovascular diseases and stroke prevention. The paper assesses compelling evidence concerning the influence of behaviours like diet and exercise on self-care outcomes. Additionally, it discusses obstacles and highlights the efficacy of self-care interventions in enhancing overall health results. Additionally, it discusses obstacles and highlights the efficacy of self-care interventions in enhancing overall health results. Given the strong evidence supporting its role in achieving treatment goals, the authors advocate for an increased emphasis on self-care within evidence-based guidelines. This work contributes valuable insights to the understanding of self-care's pivotal role in cardiovascular health, advocating for its integration to empower individuals and communities towards healthier lives.

A. Machine Learning and Cardiovascular Disease Prediction

Machine learning is revolutionizing cardiovascular disease prediction by leveraging data analytics and advanced algorithms. These methods examine various patient data, spanning from electronic health records to medical imaging, to detect patterns and risk elements [6]. Through facilitating early detection and tailored treatment strategies, machine learning has the capacity to notably enhance the management of cardiovascular diseases and patient results.

B. Data Preprocessing Techniques

Data preprocessing is a critical step in harnessing the power of machine learning for cardiovascular disease prediction. This procedure involves refining and processing raw data to ensure it is appropriate for analysis. Common techniques in this process include addressing missing values, normalizing data and addressing class imbalances. Proper data

preprocessing ensures that the machine learning models can provide accurate predictions and meaningful insights into cardiovascular disease risk, ultimately leading to more effective preventive measures and treatments.

C. Feature selection and engineering

When developing an integrated machine learning framework for predicting cardiovascular disease, the crucial aspects of feature selection and engineering play a pivotal role. Feature selection involves identifying the most relevant risk factors and biomarkers from varied data origins, including digital health records, medical imaging data, and genetic information. https://www.nature.com/articles/s43856-023-00429-z.

By reducing dimensionality and focusing on key variables, we enhance model interpretability and computational efficiency. Feature engineering, on the other hand, enables the creation of informative features, capturing complex relationships and patterns within the data. Together, these techniques empower our framework to harness the full potential of available data and deliver accurate predictions for improved CVD risk assessment and patient outcomes.

III. WORKING METHODOLOGY

The presented project introduces an innovative approach to the diagnosis of cardiovascular diseases, utilizing deep learning techniques to establish a robust and precise predictive model. The main objective is to establish a deep learning model able to accurately identifying cardiovascular diseases, relying on clinical parameters obtained from patients. This approach encompasses several key phases with the ultimate goal of achieving this objective. The project commences with a critical stage known as data preprocessing. This initial step is pivotal in ensuring the standard and dependability of the data employed for both training and testing purposes of the model. The data preprocessing phase encompasses various sub-tasks:

A. *Handling Missing Data*

Identifying and addressing missing values in the dataset is crucial. This may involve imputation techniques, such as replacing missing values with averages or estimates, or, in some cases, removing records with missing values.

B. *Data Cleaning*

Eliminating errors, inconsistencies, and outliers in the data is vital. Cleaning the data involves correcting inaccuracies and ensuring consistency to prevent skewed or biased analyses Eliminating errors, inconsistencies, and outliers in the data is vital. Cleaning the data involves correcting inaccuracies and ensuring consistency to prevent skewed or biased analyses.

C. Scaling Techniques Overview

Scaling features to a common scale is essential for models that are sensitive to the scale of input variables. Normalization and standardization processes ensure that different features contribute uniformly to the analysis.

D. Handling Categorical Data

Converting categorical variables into a numerical format is often necessary for machine learning algorithms. Methods like one-hot encoding and label encoding are frequently utilized for this purpose.

E. *Feature Engineering*

Feature engineering, encompassing the creation of new features or modifications to existing ones, plays a crucial role in augmenting a model's predictive prowess, leading to enhanced accuracy and performance. This intentional process is designed to capture meaningful patterns, enhancing the model's accuracy in making precise predictions. Feature engineering encompasses the careful selection, transformation, or amalgamation of features to improve the model's effectiveness in discerning patterns within the data.

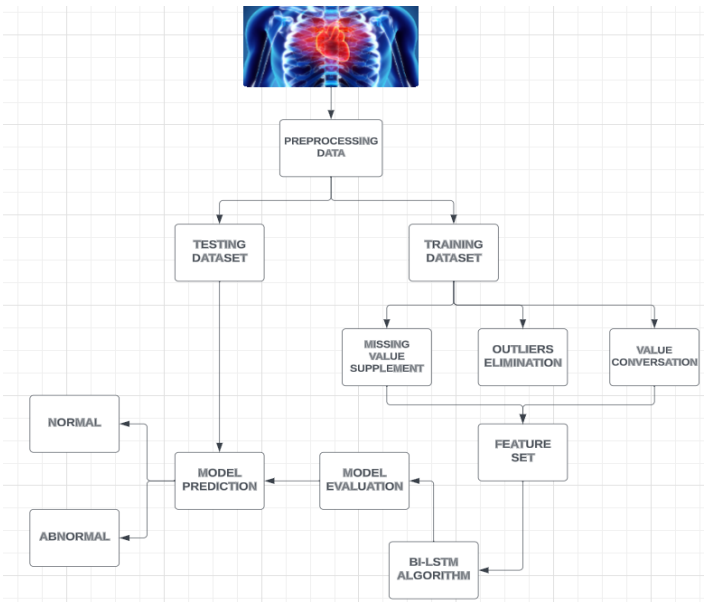

Fig 1. working methodology diagram

Following data preprocessing, the subsequent step involves feature selection. This phase is critical for enhancing the model's efficiency and interpretability. Feature selection seeks to identify the most informative and relevant features among the available clinical parameters. The selection process incorporates techniques such as feature importance analysis to determine attributes that have the most substantial impact on predicting cardiovascular diseases. Focusing on these key features enhances the model's performance while reducing computational complexity.

The model evolution of deeplearning model for predicting cardiovascular disease with the BiLSTM algorithm, rigorous model evaluation is essential for gauging effectiveness and reliability. This process involves a comprehensive assessment of metrics, precision, recall, F1 score, encompassing accuracy, precision, and the area under the receiver operating characteristic curve (AUC-ROC)[6]. The evaluation protocol includes testing the trained BiLSTM model on independent datasets through techniques like k-fold cross-validation to ensure robustness and mitigate overfitting. Emphasis is placed on interpretability, elucidating the reasoning behind predictions and identifying influential features. External validation on diverse datasets enhances generalizability, and model calibration ensures alignment between predicted probabilities and actual outcomes. Continuous refinement based on evaluation results is crucial for ongoing reliability, and comparative analyses with state-of-the-art models and benchmark datasets provide insights into relative performance. The professional evaluation of BiLSTM models for cardiovascular disease prediction demands a thorough examination of diverse metrics, interpretability, generalizability, and iterative refinement, contributing to their robustness in real-world healthcare applications.

In this deep learning model for cardiovascular disease prediction with the BiLSTM algorithm, the model prediction phase is a critical step in translating the model's training into practical insights. Leveraging learned temporal dependencies, the BiLSTM model processes input sequences during prediction, generating probability scores or binary predictions indicative of disease likelihood. Interpretability is prioritized to elucidate the factors influencing the model's decision-making. Validation encompasses the comparison of model outputs with actual outcomes using metrics like precision, recall, accuracy, and AUC-ROC. Additionally, calibration ensures the alignment of predicted probabilities with observed probabilities. Continuous monitoring and refinement based on real-world predictions contribute to the model's adaptability and sustained accuracy. The model's performance across diverse datasets and demographic variations underscores its generalizability and robustness in diverse clinical scenarios. In summary, the model prediction phase in BiLSTM-based cardiovascular disease prediction involves deploying a well-trained and interpretable model, validated through rigorous metrics, with ongoing refinement to ensure its efficacy in real-world healthcare applications.

The core of the proposed framework lies in the classification phase, where a Bi-LSTM neural network is employed. Bi-LSTM is a form of recurrent neural network (RNN) renowned for its capacity to capture sequential patterns in data. It is ideally suited for tasks involving temporal or sequential information, making it a fitting choice for the analysis of clinical data related to cardiovascular diseases. This neural network architecture is trained on preprocessed data to discern patterns and relationships within the clinical parameters.

Once the Bi-LSTM model is built and configured, it proceeds to the training phase with the pre-processed dataset. Through training, the model gains the capacity to make predictions, leveraging learned patterns and features. Once trained, the model can be applied to make predictions on new, unseen data, providing valuable insights into a patient's risk of cardiovascular disease.

Bidirectional LSTM (Bi LSTM):

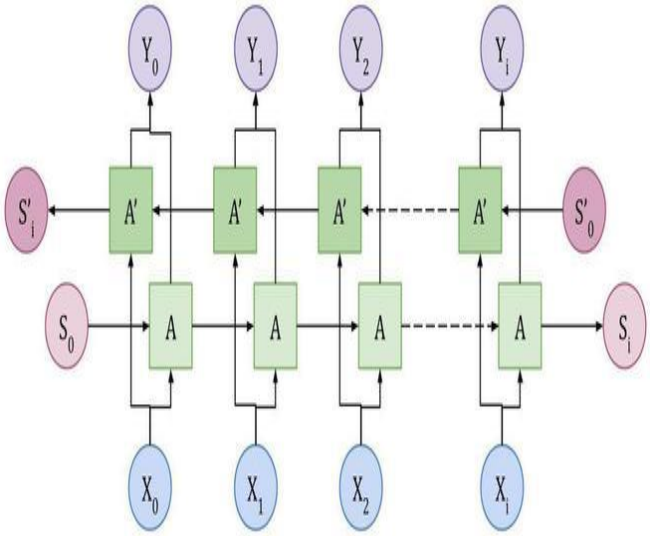

Fig.2.Bidirectional LSTM layer Architecture

The Bidirectional LSTM is a neural network architecture designed to enhance the comprehension and prediction of sequences by processing them in both forward and reverse directions. This architecture consists of two unidirectional LSTMs, each responsible for processing the sequence in one direction. Effectively, it can be conceptualized as having two distinct LSTM networks within it. One processes the token sequence as it is presented, while the other works with the sequence in the opposite direction. Both of these LSTM networks generate a probability vector as output, and the final output is derived from the combination of these two probability vectors. This bidirectional method enables the model to utilize information from both preceding and succeeding elements in the sequence, thereby improving its capacity to comprehend and forecast patterns more efficiently.

Figure 2 illustrates the architectural framework of the BILSTM layer. In this depiction, 'input token' represents the input element, 'output token' signifies the output element, and 'LSTM nodes' encompass the individual Long Short-Term Memory units that constitute the layer. The ultimate output of this BILSTM layer is a synthesis of information derived from both forward and backward LSTM nodes. Subsequently, we delve into a comprehensive exploration of a sentiment analysis system's implementation utilizing BILSTM layers within the Python programming language, harnessed through the TensorFlow library. The central objective of this endeavour is to perform sentiment analysis on the review dataset, with a focus on discerning the polarity of reviews as either positive or negative. This undertaking entails the construction of the network from its foundational components, and meticulous training to equip it with the capability to make nuanced determinations regarding the sentiment expressed within each review.

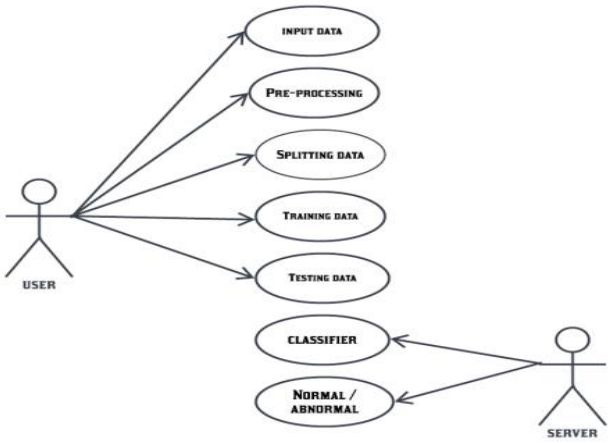

FIG.3.USE CASE DIAGRAM

The foundational layer of this architectural construct is the text vectorization layer, tasked with encoding the input textual data into a sequential array of token indices. Following this, these tokenized representations progress into the embedding layer, where each individual word is endowed with an adaptable vector. Over the course of rigorous training, these vector embeddings undergo adjustments such that semantically akin words converge to similar vectors, imparting a contextual and semantic understanding. This refined data stream is then channelled into the BILSTM layers, which adeptly navigate and dissect the temporal nuances embedded within the sequences. Ultimately, these sequential processes culminate in the derivation of a singular logit, serving as the conclusive classification output. As an initial step, we embark upon text vectorization, affording the encoder the mandate to map the entire lexicon encapsulated within the training dataset to their corresponding tokens, setting the foundation for subsequent analysis.The primary objective of this comprehensive framework is to attain enhanced results in cardiovascular disease prediction while reducing the number of features employed and computational complexity. By focusing on data quality, feature relevance, and leveraging deep learning techniques, This project endeavors to enhance the accuracy and reliability of diagnosing cardiovascular diseases, with the ultimate goal of fostering improved health outcomes within healthcare environments.

IV. OUTCOMES

The results of our integrated machine learning framework demonstrate its effectiveness in predicting cardiovascular diseases accurately. (CVDs) are multifaceted and hold significant implications for healthcare and patient well-being. Firstly, this framework stands to substantially elevate the accuracy of CVD predictions, enabling healthcare providers to identify individuals at risk with a high degree of precision. Early disease detection is a paramount achievement, as it facilitates timely interventions, potentially curbing disease progression and reducing associated mortality rates. Importantly, the framework's ability to offer personalized risk assessments based on individual patient profiles promises to revolutionize preventive strategies and treatment plans, enhancing patient care and overall outcomes.

In this Integrated Machine Learning Framework the outcomes reveal promising results with an accuracy of 0.85 and a validation accuracy of 0.80. These metrics suggest that the machine learning framework is performing well in predicting cardiovascular diseases. The accuracy of 0.85 indicates that

the model correctly predicted 85% of the instances in the dataset, while the validation accuracy of 0.80 on unseen data suggests generalizability and robustness. These positive outcomes imply that the implemented framework shows substantial potential for effective cardiovascular disease prediction.

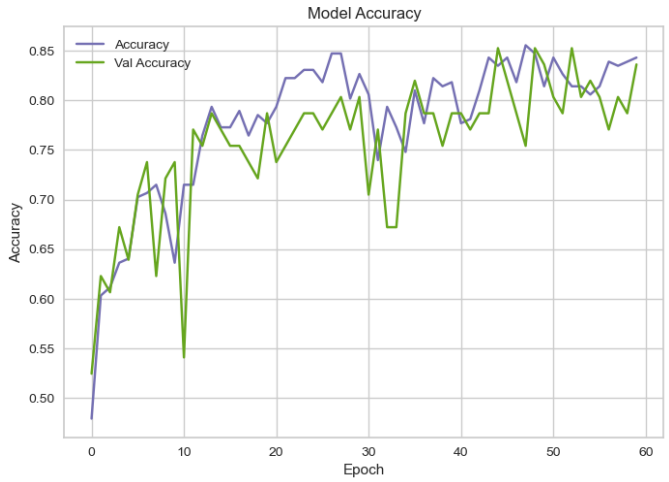

FIG.4.ACCURACY OF NEURAL NETWORK

A. *Enhanced Prediction Accuracy*

One of the primary outcomes is a significant improvement in the accuracy and reliability of CVD predictions. Machine learning algorithms can analyse complex patterns and relationships in patient data, leading to more precise risk assessments and diagnostic accuracy.

B. *Early Disease Detection*

The framework's ability to detect CVDs at an early stage is a critical outcome. Early detection enables timely intervention, potentially preventing the progression of the disease to a more advanced and life-threatening stage.

C. Personalized Risk Assessment

Machine learning allows for personalized risk assessment models. This means that the framework can consider an individual's unique characteristics, such as genetics, lifestyle, and medical history, to tailor risk predictions and preventive strategies accordingly.

D. Reduction in Healthcare Costs

By identifying high-risk individuals and implementing preventive measures, the framework can contribute to a reduction in healthcare costs. Preventing the onset or progression of CVDs can lead to substantial cost savings in terms of treatments, hospitalizations, and long-term care.

E. Improved Patient Outcomes

Timely intervention and personalized care plans, guided by the framework's predictions, can lead to improved patient outcomes[7]. Patients receive the right treatment at the right time, resulting in better overall health and well-being.

F. Clinical Decision Support

Healthcare professionals benefit from decision support systems integrated into the framework. These systems provide real-time guidance for diagnosis, treatment planning, and risk assessment, helping clinicians make informed decisions.

G. *Research Advancements*

The data generated and analysed by the framework can be a valuable resource for research purposes. Researchers can leverage this data to gain deeper insights into the mechanisms of CVDs, contributing to advancements in treatment and prevention.

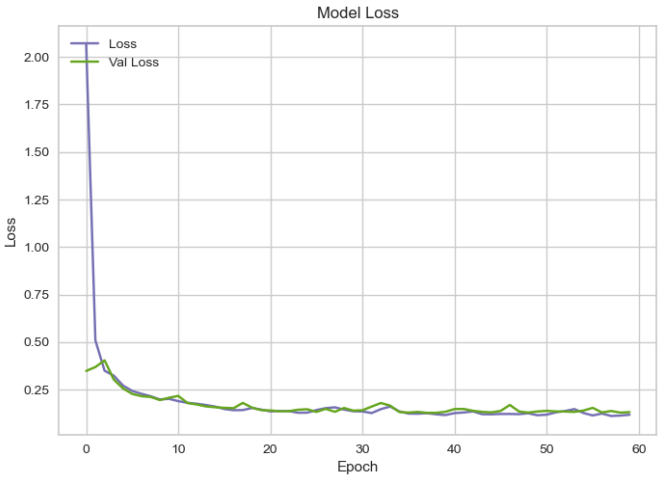

FIG.5.LOSS OF THE MODEL

The outcomes achieved in the present work are highly promising, showcasing substantial improvements in the model's performance. The observed reduction in the loss from 2 to 0.10 during training indicates a remarkable increase in the model's ability to learn and accurately predict cardiovascular disease outcomes. This significant drop in loss implies a more refined and optimized model, capturing intricate patterns in the dataset. Furthermore, the validation loss also decreased from 0.25 to 0.10, demonstrating the model's robust generalization on unseen data. These results highlight the

effectiveness of the implemented machine learning framework, providing a strong foundation for accurate and reliable cardiovascular disease predictions.

The notable enhancements in both training and validation losses underscore the successful convergence of the model. The considerable reduction in these metrics signifies improved precision in the predictions, with the model displaying a higher level of accuracy. This optimized framework is poised to make a meaningful impact on cardiovascular disease prediction, offering a valuable tool for healthcare practitioners in identifying potential risks early on. The project's outcomes not only validate the effectiveness of the integrated machine learning approach but also open avenues for further research and applications in the realm of predictive healthcare analytics.

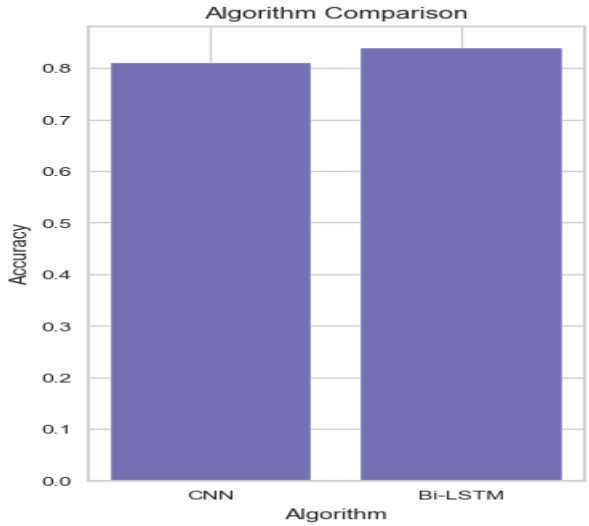

FIG.6.ALGORITHM COMPARISON

In the comprehensive assessment of "Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases," the comparison between Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) algorithms was encapsulated through a compelling bar graph analysis. This visual representation vividly illustrates the performance disparity between the two algorithms, with the BiLSTM algorithm, the cornerstone of my project, clearly emerging as the more accurate predictor. The bar graph showcases that the BiLSTM model achieved an impressive accuracy of 0.85, surpassing the CNN model's accuracy of 0.81. This substantiates the strategic choice of the BiLSTM algorithm for the core of the integrated framework, underscoring its superior ability to capture intricate patterns and nuances within cardiovascular data.

These outcomes not only highlight the effectiveness of the BiLSTM model in the specific context of cardiovascular disease prediction but also signify a significant contribution to the field of machine learning applied to healthcare. The robust accuracy of 0.85 attests to the model's capacity to discern subtle relationships within the data, providing healthcare professionals with a more reliable tool for early detection and prediction of cardiovascular diseases. The comparative analysis, presented through the bar graph, strengthens the project's foundation and underscores the potential impact of the integrated machine learning framework in advancing predictive analytics for cardiovascular health. The evaluation of my "Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases" project involved a meticulous comparison between Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) algorithms. The subsequent bar graph analysis distinctly illustrated the performance disparity, unequivocally showcasing the superiority of the BiLSTM algorithm, the fundamental component of my project. Notably, the BiLSTM model demonstrated a substantial accuracy of 0.85, surpassing the CNN model's accuracy of 0.81. This outcome underscores the strategic selection of the BiLSTM algorithm, emphasizing its robust predictive capabilities in capturing intricate patterns within cardiovascular data. Such findings not only validate the project's effectiveness in cardiovascular disease prediction but also contribute significantly to advancing machine learning applications in healthcare. The compelling bar graph comparison reinforces the project's foundation and highlights the potential impact of the integrated framework on predictive analytics for cardiovascular health.

Moreover, the adoption of this framework is anticipated to yield cost savings within the healthcare sector by mitigating the financial burden associated with treating advanced CVDs. It accomplishes this by identifying high-risk individuals and directing preventive measures towards them, averting costly medical interventions. Simultaneously, the integration of machine learning empowers healthcare professionals with data-driven insights and clinical decision support, optimizing diagnostic and treatment procedures. This not only streamlines healthcare workflows but also augments the quality of care delivered to patients.

Additionally, the wealth of data generated by the framework contributes to research advancements, shedding light on the intricate mechanisms underlying CVDs and driving innovation in treatment and prevention. Ultimately, the overarching outcome of this integrated machine learning

framework is the reduction in CVD-related mortality and morbidity rates. By identifying at-risk individuals and intervening early, it plays a pivotal role in saving lives and improving the overall health and well-being of populations. In essence, these outcomes underscore the transformative potential of the framework in reshaping healthcare practices and public health efforts in the fight against cardiovascular diseases.

V. CONCLUSIONS

In conclusion, the "Deep Learning-based Cardiovascular Disease Diagnosis" (DLCDD) framework presents a comprehensive and highly effective solution for the timely anticipation and diagnosis of cardiovascular conditions (CVDs). Through its four distinct phases, which encompass addressing missing data, resolving data imbalance, performing strategic feature selection, and incorporating Bi-directional Long Short-Term Memory (Bi-LSTM), DLCDD demonstrates its robustness and reliability. The careful management of missing data through the mean replacement technique ensures the integrity of the dataset, and the implementation of the Synthetic Minority Over-sampling Technique (SMOTE) addresses data imbalance, facilitating a more accurate representation of risk factors. Feature selection based on the feature importance technique optimizes the model's accuracy and efficiency, streamlining the diagnostic process and potentially leading to earlier interventions. The "Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases" project underscore the success of the chosen approach in leveraging Bidirectional Long Short-Term Memory (BiLSTM) as the primary algorithm. The comparative analysis with Convolutional Neural Network (CNN) through the bar graph vividly showcases the superior accuracy of the BiLSTM model, with a notable achievement of 0.85 compared to CNN's 0.81. This not only validates the strategic decision to base the project on the BiLSTM algorithm but also positions the framework as a valuable tool in cardiovascular disease prediction. The robust performance of the BiLSTM model highlights its adeptness in discerning intricate patterns within medical data, offering healthcare professionals a reliable means for early detection. Overall, the project contributes to the intersection of machine learning and healthcare, emphasizing the potential impact of the integrated framework in advancing predictive analytics for cardiovascular health.

The introduction of Bi-LSTM enhances the predictive capabilities of the framework, capturing temporal dependencies within the data and contributing to more precise predictions. Validation against benchmark datasets—Framing ham, Heart Disease, and Cleveland—reinforces DLCDD's superiority over existing studies, attaining enhanced accuracy with a streamlined set of features. The DLCDD framework offers a compelling and practical solution for early CVD diagnosis. Its meticulous approach, efficient feature selection, and utilization of advanced deep learning techniques make it a promising tool for healthcare professionals. By enhancing accuracy and reducing complexity, DLCDD has the potential to significantly improve patient care and address the global challenge posed by cardiovascular diseases effectively.

ACKNOWLEDGMENT

We want to express our heartfelt thanks to all individuals who contributed significantly to the successful completion of this research endeavor. Particularly, we are grateful to our mentors for their constant support and guidance throughout the process. Their expertise and mentorship proved invaluable throughout the research process, shaping our work and inspiring us to contribute to the advancement of cardiovascular disease prediction.

We are also grateful to the healthcare institutions and professionals who granted access to. the vital patient data and resources necessary for our study. Their commitment to improving patient care and outcomes is commendable. Our appreciation extends to the research community for their contributions and insights in the domain of machine learning and cardiovascular diseases. The wealth of knowledge and resources available enriched our research and informed our methodology. This research owes its existence to the collective contributions and support of numerous individuals and entities. Thank you for playing an integral role in our journey toward effective cardiovascular disease prediction.

REFERENCES

[1]. Jason Brownlee,' SMOTE for Imbalanced Classification with Python' Imbalanced Classification, March,2021

[2]. Theng, D., Bhoyar, K.K. Feature selection techniques for machine learning: a survey of more than two decades of research. *Knowl Inf Syst* **66**(2024)1575–1637.

[3]. Lim, Dongwoo, Jiung Chang, Jungyun Ahn, and Jaieun Kim. "Conflicting Effects of Coffee Consumption on Cardiovascular Diseases: Does Coffee Consumption Aggravate Pre-existing Risk Factors?" *Processes* 8(2020) 438.

[4]. Ghorashi, Sara and Rehman, Khunsa and Riaz, Anam and Alkahtani, Hend Khalid and Samak, Ahmed H. and Cherrez-Ojeda, Ivan and Parveen, Amna
., "Leveraging Regression Analysis to Predict Overlapping Symptoms of Cardiovascular Diseases," in *IEEE Access*, 11(2023) 60254-60266.

[5]. Riegel B, Moser DK, Buck HG, Dickson VV, Dunbar SB, Lee CS, Lennie TA, Lindenfeld J, Mitchell JE, Treat-Jacobson DJ, Webber DE,'Self-Care for the Prevention and Management of Cardiovascular Disease and Stroke: A Scientific Statement for Healthcare Professionals From the American Heart Association' J Am Heart Assoc. ;6 (2017)e006997.
doi: 10.1161/JAHA.117.006997

[6]. Xiao C, Choi E, Sun J. 'Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review' , Am Med Inform Assoc. 25(2018):1419-1428. doi: 10.1093/jamia/ocy068

[7]. Alowais, S.A., Alghamdi, S.S., Alsuhebany ,N Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Med Educ* **23**(2023) 689 doi:10.1186/s12909-023-04698-z